Machine Learning 1

Lecture 5.2 - Supervised Learning
Bayesian Linear Regression - Bayesian Model Comparison

Erik Bekkers

(Bishop 3.4)

Slide credits: Rianne van den Berg

Bayesian Model Selection

- Given L models $\{\mathcal{M}_i\}_{i=1}^{L}$ with prior belief $p(\mathcal{M}_i)$

- Update prior knowledge with observations on the data D :

$$p(\mathcal{M}_i|D) = \frac{p(D|\mathcal{M}_i)\, p(\mathcal{M}_i)}{p(D)}$$

- Predictive distribution / mixture distribution / model average:

$$p(t'|\mathbf{x}', D) = \sum_{i=1}^{L} p(t'|x', \mathcal{M}_i)\, p(\mathcal{M}_i|D)$$

- Approximation: Use most probable model for predictions

$$\mathcal{M}^* = \arg\max_{\mathcal{M}_i} p(\mathcal{M}_i|D) = \arg\max_{\mathcal{M}_i} \underbrace{p(D|\mathcal{M}_i)}\, p(\mathcal{M}_i)$$

flat prior

model selection

$$p(t'|\mathbf{x}', D, \mathcal{M}^*)$$

Bayesian Model Comparison

- Model selection

$$\mathcal{M}^* = \arg\max_{\mathcal{M}_i} p(\mathcal{M}_i|D) = \arg\max_{\mathcal{M}_i} p(D|\mathcal{M}_i)p(\mathcal{M}_i)$$

- Comparing two models $\mathcal{M}_1$ and $\mathcal{M}_2$: $\dfrac{p(\mathcal{M}_1|D)}{p(\mathcal{M}_2|D)} = \dfrac{p(D|\mathcal{M}_1)p(\mathcal{M}_1)}{p(D|\mathcal{M}_2)p(\mathcal{M}_2)}$

- When quotient of priors $\dfrac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}$ is known or close to 1, then we need

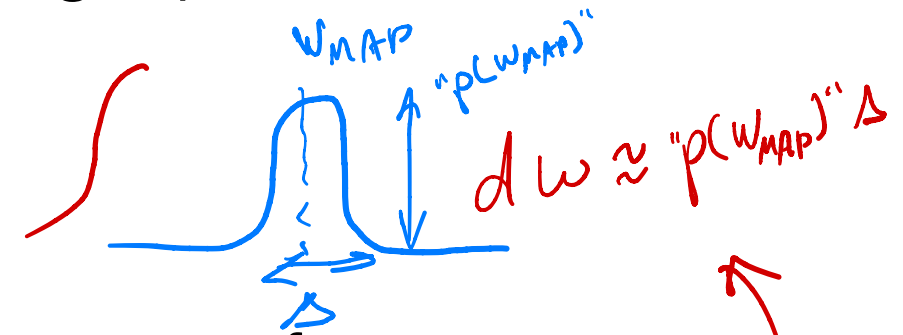$$\frac{p(D|\mathcal{M}_1)}{p(D|\mathcal{M}_2)} \qquad \textit{Bayes factor}$$

- Model evidence / marginal likelihood:

$$p(D|\mathcal{M}_i) = \int p(D|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i)\mathrm{d}\mathbf{w}$$

Approximated Model Evidence

▸ Model evidence / marginal likelihood for single parameter w

unnormalised post

$$p(D|\mathcal{M}_i) = \int p(D|w, \mathcal{M}_i)p(w|\mathcal{M}_i)\mathrm{d}w$$

w_{MAP}

$\int$ "$p(w_{MAP})$"

$dw \approx$ "$p(w_{MAP})$"Δ
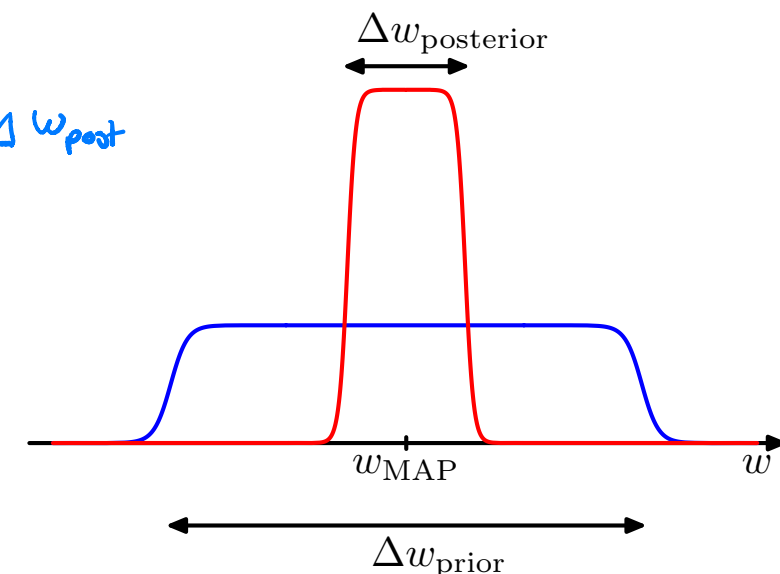
▸ Note that $p(D|M_i)$ is the normalization constant of $p(w|D, M_i)$

▸ If posterior $p(w|D, \mathcal{M}_i)$ is sharply peaked at w_{MAP} with width $\Delta w_{\mathrm{posterior}}$

$$p(w|\mathcal{M}_i) = 1/\Delta w_{\mathrm{prior}}$$

$$p(D|\mathcal{M}_i) = \int p(D|w, \mathcal{M}_i)p(w|\mathcal{M}_i)\mathrm{d}w \approx \frac{p(D|w_{MAP}\,\mathcal{M}_i)}{\Delta w_{prior}} \Delta w_{post}$$

w_{MAP}

▸ $\ln p(D|\mathcal{M}_i) \approx \ln p(D|w_{\mathrm{MAP}}, \mathcal{M}_i) + \ln \dfrac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}$

penalizes complexity

$\Delta w_{\mathrm{posterior}}$

w_{MAP} w

$\Delta w_{\mathrm{prior}}$

Figure: model evidence (Bishop 3.12)

Approximated Model Evidence

▸ $\ln p(D|\mathcal{M}_i) \approx \ln p(D|w_{\mathrm{MAP}}, \mathcal{M}_i) + \ln \dfrac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}$

▸ if $\quad \Delta w_{\mathrm{posterior}} < \Delta w_{\mathrm{prior}} \quad$ then $\quad \ln\left(\dfrac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right) < 0 \quad ^{\mathcal{M}}$

▸ M parameters: $\mathbf{w} \in \mathbb{R}^M$

$$p(D|\mathcal{M}_i) = \int p(D|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i)\mathrm{d}\mathbf{w} \approx p\left(D|w_{MAP}, M_i\right) \cdot \left(\frac{\Delta_{post}}{\Delta_{prior}}\right)^{M}$$

$$\ln p(D|M_i) \approx \ln p(D|w_{MAP}, M_i) + M \ln \frac{\Delta_{post}}{\Delta_{prior}}$$

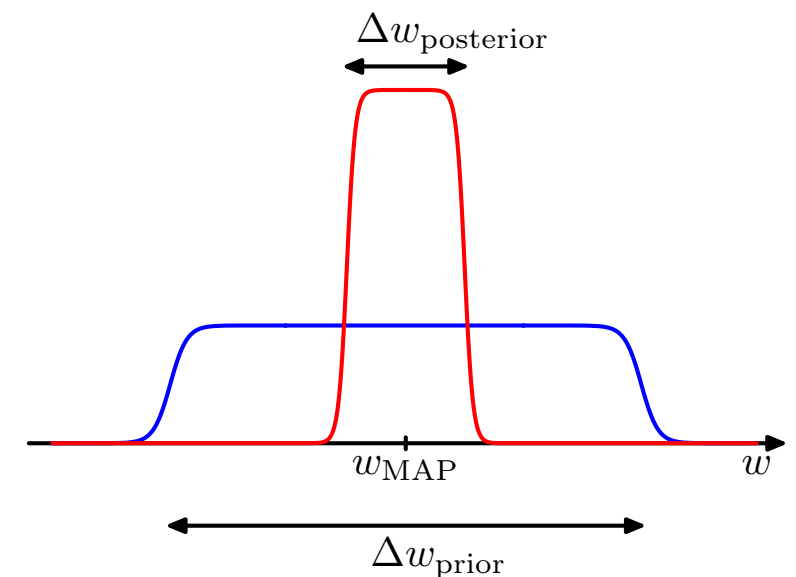▸ Model evidence favors models of medium complexity!

Figure: model evidence (Bishop 3.12)

Model evidence: medium complexity

✦ 3 models: M_1 is simplest, M_3 is most complex

✦ Generate datasets D from $p(D|M_i)$

1. sample model parameters from model prior:
$$\mathbf{w} \sim p(\mathbf{w}|M_i)$$

2. Sample dataset
$$D \sim p(D|\mathbf{w}, M_i)$$

✦ Note:
$$\int p(D|M_i)\, \mathrm{d}D = 1$$
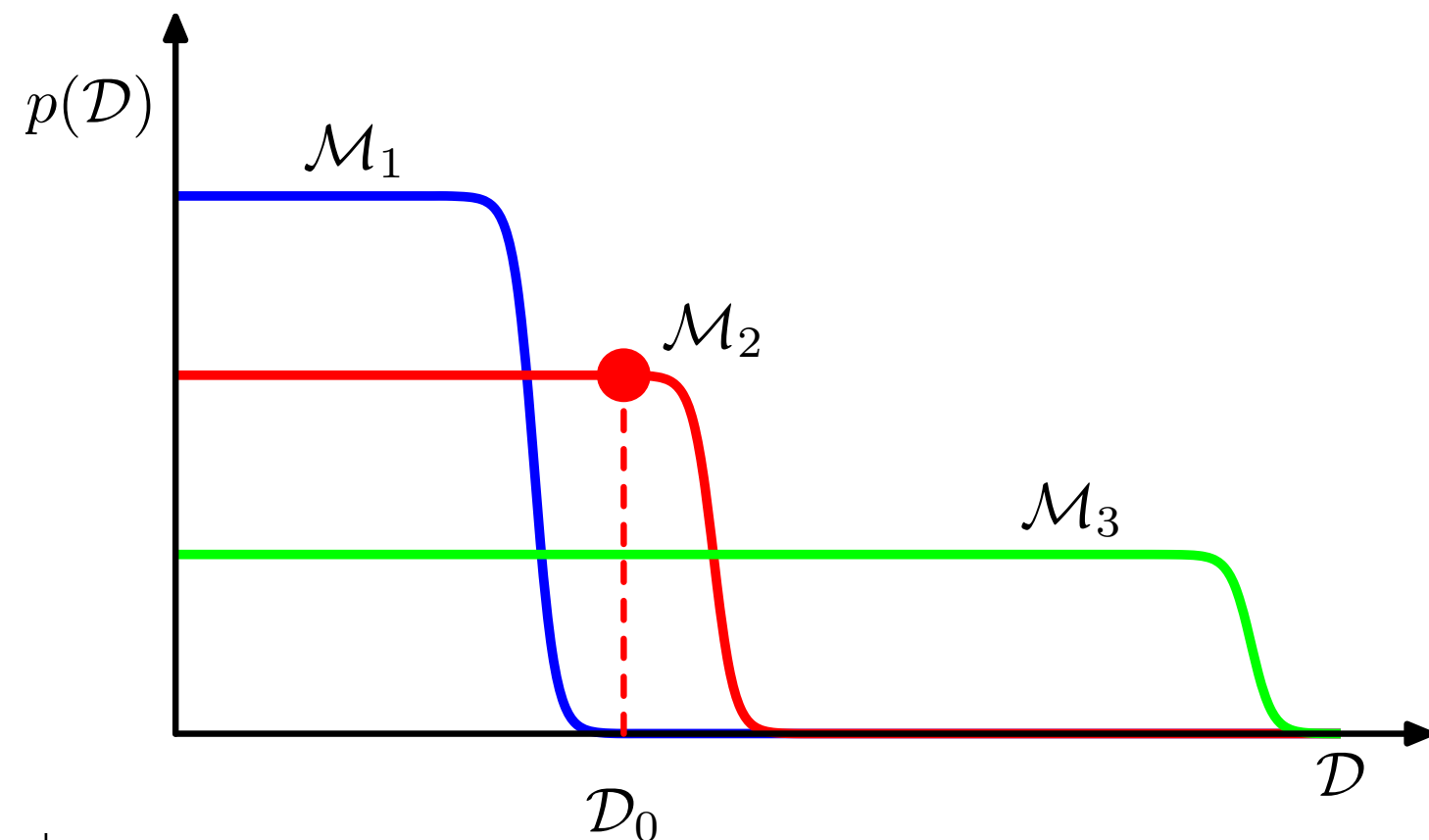

✦ dataset D_0: model M_2 has highest model evidence

Figure: model evidence (Bishop 3.12)